סיכום APML

1 בינואר 2019

אופטימיזציה

זהויות שהראנו בכיתה:

- $A\in\mathbb{R}^{m\times n}$, ו$f\left(x\right)=Ax$ כלומר טרנספורמציה לינארית אזי:

$$\frac{\partial f}{\partial x}=A$$

הוכחה:

לפי הגדרה:

$$f_{i}=\sum_{k=1}^{n}a_{i,k}x_{x}$$

ולכן: $\frac{\partial f_{i}}{\partial x_{j}}=a_{i,j}$, ולכן

$$\forall i,j:\frac{\partial f_{i}}{\partial x_{j}}=a_{i,j}\rightarrow\frac{\partial f}{\partial x}=A$$

- אם $z=y^{T}Ax$ אזי:

$$\frac{\partial z}{\partial x}=y^{T}A$$
$$\frac{\partial z}{\partial y}=x^{T}A^{T}$$

- אם A ריבועית ו $z=x^{T}Ax$

$$\frac{\partial z}{\partial x}=x^{T}\left(A^{T}+A\right)$$

- אם A ריבועית וסימטרית ו$z=x^{T}Ax$

$$\frac{\partial z}{\partial x}=2x^{T}A$$

נקודות לגראנז'

מיועד למצוא נקודות מינימום/מקסימום עם הגבלות (constraints), כלומר נרצה למצוא מינימום (למשל) ל $f(x)$ כך ש $g(x)=0$, נעשה את זה בעזרת יצירה של משוואה חדשה ואז לגזור...

אם $x\in\mathbb{R}^n$ אז המשוואה: $\frac{\partial f}{\partial x}=\lambda\frac{\partial g}{\partial x}$ היא בעצם מערכת של $n+1$ משוואות (n מהגזירה ו$g(x)=0$) נוכל לכתוב את זה כך:

$$\mathcal{L}(x,\lambda)=f(x)-\lambda g(x)$$

כדי לפתור את זה עכשיו פשוט נצטרך לגזור ולהשוואות ל0 כלומר:

$$\nabla\mathcal{L}(x,\lambda)=0$$

דוגמא:

$$f(x,y)=3x-4y$$
$$g(x,y)=x^2+y^2-1$$

כלומר למצוא מינימום של פונקציה על מעגל היחידה (הזזנו קצת את משוואת המעגל כדי שהמגבלה שלנו תיהיה $g(x,y)=0$)

$$\mathcal{L}(x,y,\lambda)=3x-4y-\lambda\left(x^2+y^2-1\right)$$

נגזור לפי כל אחד מהמשתנים ונשווה ל0:

$$\frac{\partial\mathcal{L}}{\partial x}=3-2\lambda x=0\to x=\frac{3}{2\lambda}$$
$$\frac{\partial\mathcal{L}}{\partial y}=-4-2\lambda y=0\to y=-\frac{2}{\lambda}\to x=-\frac{3}{4}y$$
$$\frac{\partial\mathcal{L}}{\partial\lambda}=x^2+y^2-1=0$$

נציב:

$$\left(\frac{3}{4}y\right)^2+y^2-1=0\to y=\pm\frac{4}{5}$$
$$x=\pm\frac{3}{5}$$

ונקבל שהנקודות הקריטיות של הפונקציה הן ב:

$$(x_1,y_1)=\left(\frac{4}{5},-\frac{3}{5}\right)$$
$$(x_2,y_2)=\left(-\frac{4}{5},\frac{3}{5}\right)$$

ועכשיו בצורה וקטורית:

נרצה למצוא את המינימום של טרסנפורמציה לינארית A כך ש היא נמצאת על כדור היחידה, כלומר $||x||_2^2 = 1$:

$$\mathcal{L}\left(x,\lambda\right)=\left(Ax\right)^{T}Ax-\lambda\left(x^{T}x-1\right)=x^{T}A^{T}Ax-\lambda\left(x^{T}x-1\right)$$

נגזור:

$$\frac{\partial\mathcal{L}}{\partial x}=2x^{T}A^{T}A-2\lambda x^{T}=0\rightarrow x^{T}\left(A^{T}A-\lambda I\right)=0$$

כלומר הו"ע עם הערך הכי קטן...

אלגוריתם EM

נדגים אותו באמצעות דוגמא, יש לנו מודל של k גאויסאנים (Mixture of Gaussians):

$$\mathbb{P}\left(x\right)=\sum_{y=1}^{k}\pi_{y}\mathcal{N}\left(x;\mu_{y},\Sigma_{y}\right)$$

π_y "המשקל" של כל גאוסיאן

μ_y התוחלת של הגאוסיאן הy

Σ_y מטריצת הcov של הגאוסיאן הy

ונרצה למצוא את הMLE שלו:

$$\hat{\theta}_{MLE}=argmax_{\theta}\left(\sum_{i=1}^{N}\log\left(\sum_{y=1}^{k}\pi_{y}\mathcal{N}\left(x;\mu_{y},\Sigma_{y}\right)\right)\right)$$

הבעיה היא הסכום שבתוך הלוג, אין לנו דרך להוציא אותו ולכן אין לנו נוסחא סגורה לפתור את המשוואה הזו.

נניח שהיינו יודעים את ההשמה של כל משתנה אינדקטור $z_{i,y}$:

$$\begin{aligned}\sum_{i=1}^{N}\log\left(\sum_{y=1}^{k}z_{i,y}\pi_{y}\mathcal{N}\left(x;\mu_{y},\Sigma_{y}\right)\right) & =\sum_{i=1}^{N}\sum_{y=1}^{k}z_{i,y}\log\left(\pi_{y}\mathcal{N}\left(x;\mu_{y},\Sigma_{y}\right)\right)\\
 & =\sum_{i=1}^{N}\sum_{y=1}^{k}z_{i,y}\left(\log\left(\pi_{y}\right)+\log\left(\mathcal{N}\left(x;\mu_{y},\Sigma_{y}\right)\right)\right)
\end{aligned}$$

כעת יכולנו להוציא את הסכום מחוץ ללוג מכיוון שרק ארגומנט אחד לא מתאפס שם.

כעת נמצא את הMLE של כל אחד מהפרמטרים:

$$
\begin{aligned}
\pi_y &= \frac{1}{N}\sum_{i=1}^{N} z_{i,y} \\
\mu_y &= \frac{\sum_{i=1}^{N} z_{i,y} x_i}{\sum_{i=1}^{N} z_{i,y}} \\
\Sigma_y &= \frac{\sum_{i=1}^{N} z_{i,y}\left(x_i-\mu_y\right)\left(x_i-\mu_y\right)^T}{\sum_{i=1}^{N} z_{i,y}}
\end{aligned}
$$

אבל אנחנו לא יודעים את $z_{i,y}$, ההסתברות שלהם היא:

$$
\mathbb{P}\left(z_{i,y}=1\right)=\mathbb{P}_\theta\left(y|x_i\right)=\frac{\mathbb{P}\left(y,x_i\right)}{\mathbb{P}\left(x_i\right)}=\frac{\pi_y\mathcal{N}\left(x_i,\mu_y,\Sigma_y\right)}{\sum_{j=1}^{k}\pi_j\mathcal{N}\left(x_i,\mu_j,\Sigma_j\right)}:=c_{i,y}
$$

ואז נוכל לעשות להם גם MLE:

$$
\mathbb{E}\left[\ell\left(S,\theta\right)\right]=\sum_{i=1}^{N}\sum_{y=1}^{k} c_{i,y}\log\left(\pi_y\mathcal{N}\left(x_i,\mu_y,\Sigma_y\right)\right)
$$

בסה"כ אלגוריתם EM הוא אלגוריתם איטראטיבי שבכל איטראציה מעדכן כלהלן:

$$
\begin{aligned}
c_{i,y} &= \frac{\pi_y\mathcal{N}\left(x_i,\mu_y,\Sigma_y\right)}{\sum_{j=1}^{k}\pi_j\mathcal{N}\left(x_i,\mu_j,\Sigma_j\right)} \\
\pi_y &= \frac{1}{N}\sum_{i=1}^{N} z_{i,y} \\
\mu_y &= \frac{\sum_{i=1}^{N} z_{i,y} x_i}{\sum_{i=1}^{N} z_{i,y}} \\
\Sigma_y &= \frac{\sum_{i=1}^{N} z_{i,y}\left(x_i-\mu_y\right)\left(x_i-\mu_y\right)^T}{\sum_{i=1}^{N} z_{i,y}}
\end{aligned}
$$

הוכחת התכנסות של EM

נרצה להראות כי $\ell\left(S,\theta^{(t)}\right)\leq\ell\left(S,\theta^{(t+1)}\right)$.
פונקציית העדכון של המשקולות היא:

$$
Q\left(\theta,\theta^t\right)=\sum_{y=1}^{k} p\left(y|x,\theta^t\right)\cdot\log\left(p\left(x,y;\theta\right)\right)
$$

נסמן $q(y) := p(y|x,\theta^t)$ נקבל:

$$\begin{aligned}
Q\left(\theta,\theta^t\right) &= \sum_{y=1}^{k} q(y)\cdot\log\left(p(x,y;\theta)\right) = \sum_{y=1}^{k} q(y)\cdot\log\left(p(x;\theta)\cdot p(x|y;\theta)\right)\\
&= \sum_{y=1}^{k} q(y)\cdot\log p(x;\theta) + \sum_{y=1}^{k} q(y)\cdot\log p(x|y;\theta)\\
&= \log p(x;\theta) + \sum_{y=1}^{k} q(y)\cdot\log p(x|y;\theta)\\
&= \ell(S,\theta) + \sum_{y=1}^{k} q(y)\cdot\log p(x|y;\theta)\\
&= \ell(S,\theta) + \sum_{y=1}^{k} q(y)\cdot\frac{\log p(x|y;\theta)}{q(y)} + \sum q(y)\cdot\log q(y)\\
&= \ell(S,\theta) - D_{KL}\left(q(y)\,||p(y|x;\theta)\right) - H\left(q(y)\right)
\end{aligned}$$

כעת נשים לב בשלב השני של כל איטריציה (M step) אנחנו מחפשים את המקסימום של D_{KL} כלומר:

$$Q\left(\theta^{(t+1)},\theta^{(t)}\right) \geq Q\left(\theta^{(t)},\theta^{(t)}\right)$$

ובסה"כ:

$$\begin{gathered}
0 \leq Q\left(\theta^{(t+1)},\theta^{(t)}\right) - Q\left(\theta^{(t)},\theta^{(t)}\right) =\\
= \ell\left(S,\theta^{(t+1)}\right) - \ell\left(S,\theta^{(t)}\right) - D_{KL}\left(q(y)\,||p\left(y|x;\theta^{(t+1)}\right)\right) + \underbrace{D_{KL}\left(q(y)\,||p\left(y|x;\theta^{(t)}\right)\right)}_{=0} - H\left(q(y)\right) + H\left(q(y)\right)\\
\leq \ell\left(S,\theta^{(t+1)}\right) - \ell\left(S,\theta^{(t)}\right)\\
\downarrow\\
\ell\left(S,\theta^{(t+1)}\right) \geq \ell\left(S,\theta^{(t)}\right)
\end{gathered}$$

הורדת רעש מתמונה

- x תמונה נקייה
- y תמונה רועשת

למה שימושית:

אם $x\sim\mathcal{N}\left(0,\sigma_x^2\right)$, $y|x\sim\mathcal{N}\left(x,\sigma^2\right)$ אזי $x|y\sim\mathcal{N}\left(\frac{\frac{1}{\sigma}y}{\frac{1}{\sigma^2}+\frac{1}{\sigma_x^2}},\frac{1}{\frac{1}{\sigma^2}+\frac{1}{\sigma_x^2}}\right)$

גאוס מרקוב:

אם $P(x,y)$ ידוע, ואנו מחפשים אלגוריתם $A(y)$ שממזער את השגיאה הריבועית הממוצעת הוא:

$$A(Y)=\mathbb{E}[x|y]$$

* שגיאה ריבועית ממוצעת: $MSE=E_{x,y}\left[||A(y)-x||^2\right]$
הוכחה:

$$\begin{aligned}MSE(A)&=\int_x\int_y p(x,y)(A(y)-x)^2\,dxdy\\&=\int_x\int_y p(y)\,p(x|y)(A(y)-x)^2\,dxdy\\&=\int_y p(y)\underbrace{\int_x p(x|y)(A(y)-x)^2\,dxdy}_{MSE_y(A)}\\&=\int_y p(y)\,MSE_y(A)\,dy\end{aligned}$$

נגזור ונשווה לאפס:

$$\begin{aligned}\frac{\partial MSE}{\partial A}=&2\int_x p(x|y)(A(y)-x)\,dx=0\\\Rightarrow&\int_x p(x|y)\,A(y)\,dx=\int_x p(x|y)\cdot xdx\\\Rightarrow&A(y)\underbrace{\int_x p(x|y)\,dx}_{1}=\mathbb{E}[x|y]\end{aligned}$$

אז הראנו שהפיתרון האופטימלי הוא $\mathbb{E}[x|y]$ אבל מה זה?
ההנחה שלנו היא ש: $y=Hx+\eta$ כך ש: $x\sim\mathcal{N}(\mu,\Sigma)$, $y|x\sim\mathcal{N}\left(Hx,\sigma^2I\right)$
נטען ש $\mathbb{E}[x|y]=\left(\Sigma^{-1}+\frac{1}{\sigma^2}H^TH\right)^{-1}\left(\Sigma^{-1}\mu+\frac{1}{\sigma^2}H^Ty\right)$
<u>הוכחה:</u>
ראשית נשים לב כי $x\sim\mathcal{N}(\mu,\Sigma)$ לכן $\mathbb{E}[x]=\underset{x}{argmax}\,(p(x))=\mu$
כמו כן, כי$\mathbb{E}[x|y]=\underset{x}{argmax}\,(p(x|y))=\underset{x}{argmax}\,(p(x,y))$, כי $x|y$ מתפלג נורמלי
אז גם (x,y) מתפלג נורמלי
אזי כל מה שנצטרך לחשב זה:

$$\begin{aligned}p(x,y)=p(x)\cdot p(y|x)&=\frac{1}{Z}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}e^{-\frac{1}{2\sigma^2}(y-Hx)^TI(y-Hx)}\\&=\frac{1}{Z}e^{-\frac{1}{2}\left((x-\mu)^T\Sigma^{-1}(x-\mu)-\frac{1}{\sigma^2}(y-Hx)^T(y-Hx)\right)}\end{aligned}$$

כאשר Z הוא גורם נירמול (שיש לנו בהתפלגות), כיוון שאנחנו מתעניינים במקסימום של x מספיק לעשות מקסימום על האקספוננט

$$\frac{\partial}{\partial x}\left(\frac{1}{2}\left((x-\mu)^T\Sigma^{-1}(x-\mu)-\frac{1}{\sigma^2}(y-Hx)^T(y-Hx)\right)\right)=(x-\mu)^T\Sigma^{-1}-\frac{1}{\sigma^2}\left(x^TH^TH-y^TH\right)=0$$

נפתור:

$$x^T\left(\Sigma^{-1}+\frac{1}{\sigma^2}H^TH\right)=\mu^T\Sigma^{-1}+\frac{1}{\sigma^2}y^TH$$

$$x^T=\mu^T\Sigma^{-1}+\frac{1}{\sigma^2}y^TH\left(\Sigma^{-1}+\frac{1}{\sigma^2}H^TH\right)^{-1}$$

$$x=\left(\Sigma^{-1}+\frac{1}{\sigma^2}H^TH\right)^{-1}\mu^T\Sigma^{-1}+\frac{1}{\sigma^2}y^TH$$

כעת אנו מתעיינים רק ב$H=I$ ונקבל:

$$x=\left(\Sigma^{-1}+\frac{1}{\sigma^2}I\right)^{-1}\left(\mu^T\Sigma^{-1}+\frac{1}{\sigma^2}y^T\right)$$

זיהוי חלקי משפט

סימונים:

- $X_{1:n}$ המילים במשפט
- $Y_{1:n}$ התיוגים POS של המשפט

כאשר n אורך המשפט.

מניחים מרקוביות (לאו דווקא נכון), כלומר (שרשרת מסדר ראשון):

$$P\left(Y_i|Y_1,\ldots,Y_{i-1}\right)=P\left(Y_i|Y_{i-1}\right)$$

במילים - כדי לתת ציון לשרשרת תיוגים מספיק לתת ציון לכל זוג תיוגים (זו שרשרת מסדר ראשון, אפשר להרחיב לסדר שני)

מודל HMM

- $Y_{1:n}$ שרשרת מרקובית (כמו שכתבנו)
- המ"מ X_i בת"ל בכל דבר חוץ מY_i, כלומר אם נדע את Y_i זה לא ממש משנה איזו מילה בדיוק נמצאת בX_i כדי לחזות את שאר המילים במשפט, ההנחה הזו בבירור לא מתקיימת.

בנוסף יש לנו 2 סוגים של פרמטרים:

- הסתברויות מעבר: $t\left(y,y'\right)=\mathbb{P}\left(Y_i=y|Y_{i-1}=y'\right)$, וכן $\sum_y t\left(y,y'\right)=1$
- הסתברות emission: לכל y ערך אפשרי של Y וw של X:

$$e\left(w,y\right)=\mathbb{P}\left(X_i=w|Y_i=y\right)$$

וכן $\sum_w e\left(w,y\right)=1$

נשאר לנו ללמוד את e,t ולהבין איך לעשות פרדיקציה כמו שצריך.

הסקה - viterbi

איך נעשה inference?

$$y^*=\underset{y_{1:n}}{argmax}\mathbb{P}\left(y_{1:n}|x_{1:n}\right)$$

נשים לב ש:

$$\mathbb{P}\left(y_{1:n}|x_{1:n}\right)=\frac{\mathbb{P}\left(y_{1:n}\right)\cdot\mathbb{P}\left(x_{1:n}|y_{1:n}\right)}{\mathbb{P}\left(x_{1:n}\right)}$$

ולכן:

$$y^*=\underset{y_{1:n}}{argmax}\mathbb{P}\left(y_{1:n}|x_{1:n}\right)=\underset{y_{1:n}}{argmax}\mathbb{P}\left(y_{1:n}\right)\cdot\mathbb{P}\left(x_{1:n}|y_{1:n}\right)$$

זה בדיוק הפירוק שעשינו קודם ל transition ו emission כלומר :

$$y^*=\underset{y_{1:n}}{argmax}t\left(y,y^{'}\right)\cdot e\left(w,y\right)$$

(כי נזכר ש: $t\left(y,y'\right)=\mathbb{P}\left(Y_i=y|Y_{i-1}=y'\right)$, $e\left(w,y\right)=\mathbb{P}\left(X_i=w|Y_i=y\right)$)
אפשר לפתור את זה בתכנון דינמי - viterbi
כאמור הבעיה הכללית שלנו היא למצוא את:

$$\pi\left(t,j\right)=\underset{y_1\cdots y_{t-1}}{max}P\left(y_{1:(t-1)},y_t=j,x_{1:t}\right)$$

כלומר, הרצף עם ההסתברות הכי גדולה של $y_1,\ldots,y_t$ שנגמר בj כל זה בהינתן הרצף $x_{1:t}$

ואפשר לפרק את זה לתתי בעיות:

$$\pi\left(t,j\right)=\underset{j'\in S}{max}\left\{\pi\left(t-1,j'\right)\cdot t\left(j,j'\right)\cdot e\left(x_t,j'\right)\right\}$$

תנאי התחלה:

$$\pi\left(1,j\right)=t\left(j,START\right)\cdot e\left(x_1,j\right)$$

כאשר בסוף מה שמעניין אותנו הוא:

$$max_{i\in S}\pi\left(n,i\right)$$

כלומר כל המילים, ומה ההשמות עם הסיכוי הכי גבוה שלהם.. (S זו קבוצת התיוגים האפשריים למשפט באורך n, כלומר אם k זה מספר אופציות התיוג האפשריות $|S|=n^k$)

למידה

כאמור ההתפלגות המשותפת של $Y_{1:n},X_{1:n}$ מעניינת אותנו, למשפט בודד ההסתברות היא:

$$Pr\left(Y_{1:n},X_{1:m}\right)=Pr\left(Y_1\right)\cdot\prod_{i=2}^{n}Pr\left(Y_i|Y_{i-1}\right)\cdot Pr\left(X_i|Y_i\right)$$

כלומר:

$$Pr\left(Y_{1:n},X_{1:m}\right)=\prod_{i=1}^{n}t\left(y,y^{'}\right)\cdot e\left(x,y\right)$$

אבל יש לנו גם אילוצים:

- כל המשתנים אי שליליים
- $\forall y'\in T\sum_{y\in T}Pr\left(y|y'\right)=1\wedge\forall y\in T\sum_{x\in S}Pr\left(x|y\right)=1$

נסתכל על הlog likelihood של המשוואה הראשונה (ללא האילוצים כרגע):

$$LL=\sum_{k=1}^{n}\sum_{i=1}^{n_k}\log Pr\left(y_i^{(k)}|y_{i-1}^{(k)}\right)+\log Pr\left(x_i^{(k)}|y_i^{(k)}\right)$$

לא בטוח שצריך את האילוצים, נשתמש בשיטת לגראנז'

$$\mathcal{L}=LL-\lambda_1\left(\sum_{y\in T}Pr\left(y|y'\right)-1\right)-\lambda_2\left(\sum_{x\in S}Pr\left(x|y\right)-1\right)$$

נגזור נשווה לאפס

$$\frac{\partial}{\partial\left(Pr\left(y|y'\right)\right)}=n_{ij}\cdot\frac{1}{Pr\left(y|y'\right)}-\lambda_1=0\Rightarrow Pr\left(y|y'\right)=\frac{n_{i,j}}{\lambda_1}\to Pr\left(y|y'\right)=\frac{n_{i,j}}{\sum_k n_{j,k}}$$

- n_i מספר הפעמיים שi התחיל משפט
- $n_{i,j}$ מספר הפעמיים שj הופיע אחרי i

- $n_{x,i}$ מספר הפעמיים שx קיבל את התיוג i
- T אוסף המילים
- S אוסף התיוגים

נחזור על התהליך לכולם ונקבל:

$$\hat{q}\left(i\right)=\frac{n_{i}}{\sum_{i}n_{i}}$$
$$\hat{e}\left(w,i\right)=\frac{n_{w,i}}{\sum_{w}n_{w,i}}$$
$$\hat{t}\left(x,i\right)=\frac{n_{x,i}}{\sum_{x}n_{x,i}}$$

אחת הבעיות שיש לנו כאן זה דוגמאות שלא ראינו בתהליך האימון יקבלו 0 מיד...

רשתות קונבולוציה

פונקציות:

- אקטיבציה של נוירונים באמצע הרשת:

$$ReLu\left(z\right)=max\left(0,z\right)$$
$$Sigmoid\left(z\right)=\frac{1}{1+e^{-z}}$$

- מחיר (לשכבה האחרונה של הרשת):
  - ריבועי square loss נשתמש לבעיות רגרסיה

$$L\left\{ w_{ij}^{(l)}\right\} =\sum_{m=1}^{M}||\hat{y}^{m}-y^{m}||_{2}^{2}$$

  - לבעיות סיווג (classification) נשתמש בsoftmax כי הוא דואג שהסכום יהיה 1 (נראה לי)

$$L_{s}\left(\left\{ w_{ij}^{(l)}\right\} \right)=-\sum_{m=1}^{M}y^{m}\cdot\log\hat{y}^{m}$$

חוץ מאקטיביציה עוד שני פרמטרים חשובים לרשתות:

– מומנטום ־ במקום שיהיה רק קצב למידה (lr) $x+=-lr\cdot dx$, נשתמש גם במה שקרה קודם כלומר

$$v = mu \cdot v - lr \cdot dx$$
$$x+ = v$$

אפשר לשחק עם זה עוד (למשל מסדר שני כמו בADAM)

– ההעפה של נוריונים ־ dropout. כדי שהרשת תלמד להכליל טוב יותר

* בתהליך הלמידה בהסתברות p לא נשתמש נאפס נוריון מסוים (במעבר הבא ברשת נדגום שוב לא נאבד את המשקולות שלו)
* בתהליך הserving נכפיל את התוצאה שלנו בp (במקרה הלינארי זה בדיוק אותו דבר ובסה"כ זה קירוב מספיק טוב במקום להריץ מלא פעמיים)
* דרך אחרת לעשות את זה זה תוך כדי האימון לחלק בp את כל המשקולות (ועדיין לאפס בהסתברות p)

גודל התוצאה של קונבולוציה:

נסמן:

- N גודל ה"תמונה" ($N\times N$)
  - c מספר הערוצים בתמונה (כלומר בעצם התמונה האי $N\times N\times c$)
  - F גודל הקרנל
  - k מספר השכבות בקרנל (כל שכבה היא בעצם c מימדית)
  - p padding כלומר כמה אפסים מוסיפים
  - s stride כלומר כמה מדלגים בין הפעלת קונבולוציה אחת לאחרת

אזי גודל התוצאה שלנו יהיה:

$$O = \frac{N + 2 \cdot p - F}{s} + 1$$

ובסה"כ $O\times O\times k$ (נשים לב שמספר הערוצים השתנה!)

נשים לב שכדי שהקונבולוציה תיהיה חוקית צריך שO יהיה מספר שלם

כמות פרמטרים:

לכל שכבה בקרנל יש לנו: $F^2\cdot c+1$ ה־1 מגיע מהbias

לכן בסה"כ $k\left(F^2\cdot c+1\right)$

נשים לב שזה ממש הגיוני לעשות קונבולציה 1×1 כי זה יוריד את מספר הערוצים

הליכה אחורה - Back Propagation

ראשית נזכר בכלל השרשרת:

$$(f \circ g)^{'} = \left(f^{'} \circ g\right) \cdot g^{'}$$

נניח שיש לנו את הפונקציה: $f(x) = x^2 + 5x^3$
אפשר לחשוב על זה כך:

$$\begin{aligned} h(x) &= x^2 \\ q(x) &= x^3 \\ s(x) &= 5x \\ p(x,y) &= x+y \\ f(x) &= p(h,q) \end{aligned}$$

וכעת אם נרצה לגזור $\frac{\partial f}{\partial x}$ נעשה זאת כך:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial p}\left(\frac{\partial p}{\partial h} \cdot \frac{\partial h}{\partial x} + \frac{\partial p}{\partial s} \cdot \frac{\partial s}{\partial q} \cdot \frac{\partial q}{\partial x}\right) = 1 \cdot 2x + 1 \cdot 5 \cdot 3x^2 = 2x + 15x^2$$

כפי שהיינו מצפים..

הבעיה בכך זה שזה כמובן יכול להיות אקספוננציאלי, מה שנעשה זה כשנעבור קדימה ברשת נחשב לכל קודקוד את הנגזרת החלקית שלו (בדוגמא h, q, s, p) ואז נלך אחורה ונחשב את המכפלה...

קלאסטרינג

שיטת silhoutte

לכל נקודה נגדיר: מרחק ממוצע בתוך הקלאסטר a_i,

מרחק ממוצא לקלאסטר השכן: $b_i = \min_{j, x_i \notin C_j} \frac{1}{|C_j|} \sum_{x \in C_j} ||x - x_i||$ במילים - נחשב את ממוצע המרחק בין הנקודה שלנו x_i לכל אחד מהקלאסטרים (חוץ משל הנקודה כמובן) C_j ונבחר את b_i להיות המינימום.

נגדיר את המרחק

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

ככל שממוצע s_i לקלאסטר גדול יותר כך הוא הדוק ומרוחק יותר $\Leftarrow$ יותר טוב

פירוק ספקטרלי

מטריצת לפליסאן

W מטריצת השכנויות שלנו (נשיג אותה ע"י סף או קרנל או כל דרך אחרת)

D מטריצת דרגות, כלומר: $D_{i,j}=\begin{cases}0 & i\neq j\\ \sum_k w_{ik} & i=j\end{cases}$

$L=D-W$

נטען שL סימטרית, כי D אלכסונית וW סימטרית.

נרצה להראות שL היא PSD, יהי x וקטור כלשהו

$$\begin{aligned}x^TLx=x^TDx-x^TLx&=\sum_i x_i^2D_{ii}-\sum_{i,j}x_ix_jW_{ij}\\&=\frac{1}{2}\left(\sum_i\left(\sum_j W_{ij}\right)x_i^2-2\sum_{i,j}x_ix_jW_{i,j}+\sum_j\left(\sum_i W_{ij}\right)x_j^2\right)\\&=\frac{1}{2}\sum_{i,j}W_{i,j}\left(x_i-x_j\right)^2\geq 0\end{aligned}$$

וזהו תנאי מספיק להיותה של L PSD

אחת התכונות של PSD זה שכל הע"ע שלה חיובים.

הריבוי של הע"ע 0 יהיה מספר רכיבי הקשירות בגרף.